LE PETIT LAROUSSE ILLUSTRÉ DE 1905 EN LIGNE : SECRETS DE FABRICATION ET PRÉSENTATION

Helene Manuelian, Audrey Bruscand, Nicole Cholewka, Anne-Marie Hetzel

LE *PETIT LAROUSSE ILLUSTRÉ* DE 1905 EN LIGNE : SECRETS DE FABRICATION ET PRÉSENTATION

Résumé : Dans cet article sont présentés les procédés utilisés pour informatiser le Petit Larousse Illustré *de 1905 et en proposer une version interrogeable en ligne. Dans un premier temps, nous explicitons les différentes étapes de la fabrication et les choix théoriques et informatiques qui ont été faits pour parvenir à l'informatisation, puis nous présentons la base de données elle-même et les différentes possibilités d'interrogation offertes.*

1 INTRODUCTION : PRÉSENTATION DU PROJET

1. 1 Naissance du projet

Le projet d'informatisation du *Petit Larousse* est né au sein de l'UMR Métadif en 2004, de la constatation qu'il devenait de plus en plus nécessaire à la métalexicographie de disposer de dictionnaires anciens informatisés. En effet, les dictionnaires contemporains possédaient alors tous (ou presque) leur version informatisée, et avaient démontré (Pruvost, 2000) l'intérêt de la consultation électronique. On pouvait enfin revenir à une consultation analogique des dictionnaires, mais aussi procéder à des interrogations étonnantes, impossibles ou très longues à réaliser avec un dictionnaire papier.
Aussi, sous la direction de Jean Pruvost, a commencé un travail d'analyse puis d'informatisation de l'un des premiers dictionnaires illustrés de l'histoire lexicographique française, le *Petit Larousse Illustré* né en 1905.

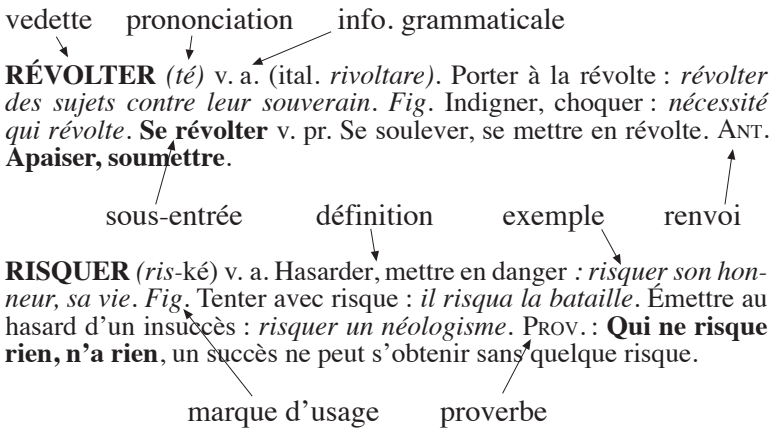
1.2 Buts et contraintes

Les buts de l'informatisation du *Petit Larousse* de 1905 étaient multiples. Ils étaient tout d'abord scientifiques. Le *Petit Larousse* regorge d'exemples qu'il était souhaitable de pouvoir interroger séparément du reste du texte. Il est aussi tout à fait représentatif de l'idéologie républicaine véhiculée par la maison Larousse, et l'interroger informatiquement représentait une chance de révéler toutes les facettes de cette idéologie. En d'autres termes, force raisons scientifiques étaient présentes derrière le projet.
D'autres buts corollaires aux buts scientifiques sont vite apparus. En 2004, le fac-similé du premier *Petit Larousse* publié pour le centenaire (par la force

des choses) n'existait pas encore. Le dictionnaire, dans sa version papier, était donc un objet rare, fragile voire en mauvais état. Sa version informatique permettait donc sa conservation et sa diffusion, bien mieux qu'une édition papier.
Les contraintes que nous avions alors étaient les suivantes : - conserver la partie la plus caractéristique du dictionnaire : les images ; - permettre une consultation plein texte et une consultation experte du dictionnaire ; - en assurer une consultation universelle et pérenne.
Il nous fallait donc numériser le texte et les images, en faire une annotation sémantique (pour permettre l'interrogation experte) tout en respectant les standards de balisage informatique (pour permettre la consultation la plus large possible).

1.3 Solutions adoptées

Tout d'abord, afin de baliser sémantiquement les articles du dictionnaire, il a fallu en réaliser l'analyse détaillée. C'est ainsi que nous avons pu découper les articles en différents champs : la vedette, la prononciation, l'étymologie, les informations grammaticales, les notes de conjugaison, les marques d'usage (de style, de registre, de domaine…), les renvois (antonymiques, synonymiques, et analogiques), les sous-entrées (forme pronominales de verbes, nominalisations, expressions…), les proverbes et locutions proverbiales, les exemples, les définitions et les définitions encyclopédiques.
Il est alors apparu très clairement que la forme typographique du texte correspondait précisément au découpage réalisé. Par ailleurs, la structure des articles du *Petit Larousse Illustré* nous a paru très régulière. Ces deux constats étant faits, nous avons acquis la certitude que nous pouvions utiliser la forme du texte et la structure des articles pour baliser sémantiquement le dictionnaire, comme nous le constatons dans la figure ci-dessous[1] :



1. Nous ne faisons pas apparaître tous les éléments dans les légendes pour conserver la lisibilité du schéma.

Il a fallu ensuite choisir un format informatique pour la base lexicale que nous souhaitions créer. XML s'est imposé, puisqu'il s'agit d'un format standard pour toutes les bases lexicales. À partir de l'analyse ayant permis le découpage des articles, nous avons ainsi pu puiser dans la TEI P5 (la proposition 5 de la Text Encoding Initiative) les balises XML nécessaires à l'annotation du texte. La TEI fournit tous les standards de balisage des ressources textuelles, et contient entre autres, un chapitre régulièrement révisé sur l'annotation des dictionnaires informatisés (Ide et Véronis, 1994). Il est arrivé que nous ayons à typer les balises fournies par la TEI pour les préciser, pour les désambiguïser, mais dans l'ensemble, notre travail prouve que la TEI est assez souple et assez précise à la fois pour annoter un dictionnaire informatisé.
Dans cet article, nous allons présenter les différentes étapes de la fabrication du *Petit Larousse Illustré* de 1905 en ligne et montrer comment nous avons respecté les contraintes et les choix opérés au départ du projet, puis nous présenterons les différentes possibilités d'interrogation de la base.

2 ÉTAPES DE LA FABRICATION

L'annotation du texte du dictionnaire s'est déroulée en trois phases. Antérieurement aux phases de balisage, s'est déroulée une période de numérisation (scanner et reconnaissance optique de caractères) et de relecture des fichiers issus de la numérisation. Nous ne nous étendrons pas plus avant sur cette première étape du travail, dans la mesure où elle ne présente pas d'intérêt scientifique et n'a pas apporté de renseignement particulier sur le texte traité. Disons simplement que cette phase a permis l'obtention de fichiers HTML contenant toutes les indications de mise en forme du texte (gras, italiques, petites majuscules, sauts de ligne, sauts et numéros de page), et qu'à partir de cette mise en forme, des programmes écrits en Python ont permis l'annotation sémantique et structurelle du texte.
En revanche, nous allons décrire plus précisément les trois passes d'annotation réalisées pour arriver à la version interrogeable du dictionnaire, parce qu'elles ont permis non seulement le balisage du texte mais aussi quelques découvertes surprenantes sur la mise en forme et la structuration des articles, ainsi que sur le contenu du dictionnaire. L'ordre dans lequel les programmes ont été appliqués n'est pas le fruit du hasard, mais celui d'une analyse précise de la forme et de la structure des définitions. Il n'est pas utile de la restituer ici, mais cette analyse a permis que le résultat de chacune des étapes soit utilisé dans l'étape suivante.

2.1 Première passe d'annotation

La première passe d'annotation a permis de repérer automatiquement les éléments suivants : les vedettes, les indications de prononciation, les indications grammaticales, et les marques d'usage. Ceci a été fait de la façon suivante :

2.1.1 Repérage de la vedette (balise<orth>)

La vedette est l'élément le plus facile à repérer. Il s'agit du premier mot de l'article, il est en gras et en majuscule. La plupart du temps, ces éléments

sont suffisants pour la pose de la balise <orth> qui repère la vedette dans la TEI. Pourtant, nous trouvons des variantes que nous nous devons d'annoter aussi. Ainsi, il existe soit des vedettes indiquant une variante orthographique du mot, soit une flexion du mot donné en entrée, comme dans les exemples suivants :

CUILLER ou **CUILLÈRE** *(kui, ll* mll.*, è-re)*
CUEILLEUR, EUSE (*keu, ll* mll., *eu-ze)*
CUISINIER (*zi-ni-é*), **ÈRE**

Nous sommes donc confrontés à deux orthographes possibles, séparées par la conjonction « ou » ; il arrive même que certains articles présentent trois à quatre formes orthographiques possibles en entrées, séparées par des virgules puis par la conjonction « ou ».

Est parfois offerte la forme fléchie d'un nom ou d'un adjectif, séparée du lemme soit par une virgule, soit par une indication de prononciation.

Il existe même des cas où les deux possibilités (variante orthographique et indication de flexion) sont cumulées, ce qui donne des vedettes très complexes.

Nos programmes repèrent aussi ces vedettes complexes, et récupèrent les formes fléchies complètes (féminins et pluriels), de façon à permettre les interrogations sur les formes féminines et plurielles des entrées du dictionnaire. Ainsi, les entrées citées précédemment seront-elles balisées de la façon suivante :

```
<form><orth>CUEILLEUR</orth></form>, <form type="infl"><orth
type="partial" mot_entier="cueilleuse">EUSE</orth>
<pron>(keu, ll mll., eu-ze)</pron></form>
<form><orth>CUILLER</orth></form> ou <form><orth>CUILLÈRE</
orth> <pron>(kui, ll mll., è-re)</pron></form>
<form><orth>CUISINIER</orth> <pron>(zi-ni-é)</
pron></form>, <form type="infl"><orth type="partial"
mot_entier="cuisinière">ÈRE</orth></form>
```

Les variantes orthographiques sont balisées comme la vedette présentée en vedette « principale », avec une balise <orth>. Les formes fléchies, en revanche, portent des types. Sur la balise <form> (qui englobe la balise de la vedette et celle de la prononciation), on trouve un type appelé « infl » pour « inflexion » (le terme anglais pour *flexion*), qui nous permet d'indiquer que la vedette est présentée sous plusieurs formes appartenant à son paradigme de flexion. La balise de la vedette, <orth> est elle typée avec l'indication « partial » (puisqu'elle est donnée sous une forme partielle) et nous y avons ajouté l'attribut « mot_entier » qui a pour valeur la forme entière du féminin ou du pluriel. Ces formes entières sont calculées automatiquement par le programme (sur la base des règles du français). Cette indication de la forme complète du mot fléchi ne figure bien entendu pas dans les recommandations de la TEI mais elle nous a semblé indispensable pour tout utilisateur souhaitant faire une recherche sur une forme féminine ou plurielle (cf. partie 3).

2.1.2 Repérage des indications de prononciation (balise <pron>)

Le programme qui balise les vedettes balise aussi les indications de prononciation. Ce programme-ci est moins complexe. En effet, les prononciations sont systématiquement entre parenthèses et en italique derrière une balise <orth>. Si la vedette est bien repérée, la prononciation l'est donc aussi. Il a seulement fallu intégrer une condition supplémentaire dans le programme de balisage permettant de tenir compte de l'indication « *ll* ml. » pour « l mouillé », puisque cette dernière contient des caractères qui ne sont pas en italiques (cf. l'exemple de l'entrée *cueilleur* donné précédemment). L'ensemble <orth> et <pron> est encapsulé dans une balise <form> puisqu'il s'agit des deux types d'indications que l'on peut trouver sur la forme d'une unité lexicale.

2.1.3 Repérage des indications grammaticales (balise <gramGrp>)

Les indications grammaticales sont toujours données sous forme d'abréviations. Elles apparaissent la plupart du temps derrière la prononciation ou la vedette, mais parfois, on les retrouve en milieu d'article (pour indiquer une sous-entrée). Le programme qui balise les indications grammaticales repère donc les abréviations qui sont données comme une liste fermée par les relectrices (la liste d'abréviation donnée en introduction du dictionnaire étant très largement incomplète). L'ensemble des indications grammaticales est inclus dans une balise <gramGrp>, les indications de catégorie grammaticale sont dans une balise <pos> (pour *part of speech*), les indications de genre dans une balise <gen> et de nombre également dans une balise <number>.

2.1.4 Repérage des marques d'usage

Les marques d'usage sont assurément variées dans le *Petit Larousse Illustré* de 1905. La typologie qu'en dresse la TEI nous semble tout à fait adaptée, c'est pourquoi nous l'avons adoptée dans l'annotation de notre dictionnaire. Par ailleurs, ces marques sont, comme les marques grammaticales, toujours abrégées de la même façon dans le dictionnaire. C'est pourquoi nous avons pu dans le programme qui les a balisées automatiquement en donner la liste complète. Bien entendu, de la même manière que pour les abréviations grammaticales, la liste fournie dans le dictionnaire n'était pas complète et nous avons dû la compléter lors des relectures. Toutes les marques d'usage sont repérées avec la balise <usg> portant un attribut « type » dont la valeur sera l'une des suivantes : **- plev**. pour « preference level », qui marque la préférence de l'auteur pour un usage ou une indication d'emploi très générale (généralement, principalement, etc.) ; **- dom**., qui indique une marque de domaine (médecine, science, littérature, etc.) ; **- style** indiquant une marque de style (figuré, propre, etc.) ; **- reg**., signalant une marque de registre (familier, populaire, etc.) ; **- time,** indiquant une marque sur l'époque d'utilisation du mot vedette (aujourd'hui, autrefois, etc.).
À la fin de la première passe d'annotation, les trois entrées qui nous ont servi d'exemple dans cette section sont balisées de la manière suivante :

```
<EntryFree><form><orth>CUEILLEUR</orth></
form>, <form type="infl"><orth type="partial" mot_
entier="cueilleuse">EUSE</orth> <pron>(keu, ll mll.,
eu-ze)</pron></form> <gramGrp><pos>n.</pos></gramGrp>
Celui celle qui cueille. <usg type="plev">Peu us.</usg>)</
EntryFree>
<EntryFree><form><orth>CUILLER</orth></form> ou
<form><orth>CUILLÈRE</orth> <pron>(kui, ll mll., è-re)</
pron></form> <gramGrp><pos>n.</pos> <gen>f.</gen></
gramGrp> (lat. <i>cochleare ;</i> de <i>cochlea</i>,
coquille). Ustensile de table, composé d'un manche et
d'une partie creuse pour puiser les aliments liquides ou
peu consistants. <i>Cuillère à pot</i>, grande cuillère de
cuisine. Ustensile servant à puiser les métaux en fusion.</
EntryFree>
<EntryFree><form><orth>CUISINIER</orth> <pron>(zi-ni-é)</
pron></form>, <form type="infl"><orth type="partial" mot_
entier="cuisinière">ÈRE</orth></form> <gramGrp><pos>n.</
pos></gramGrp> Qui fait la cuisine. <gramGrp><pos>N.</pos>
<gen>f.</gen></gramGrp> Appareil en fonte ou en tôle, muni
d&#39;un ou de deux foyers, et à l&#39;aide duquel on peut
faire cuire les aliments, tout en chauffant un appartement.
Sorte de rôtissoire, de coquille, destinée au grillage des
viandes.</EntryFree>
```

2.2 DEUXIÈME PASSE D'ANNOTATION

La deuxième passe d'annotation permet d'annoter les notes étymologiques, les renvois, les exemples, et les proverbes. Cette passe utilise les résultats de la précédente, dans la mesure où elle a supprimé des marques typographiques communes qui auraient pu prêter à confusion (par exemple, les étymologies sont entre parenthèses et contiennent des italiques, comme les prononciations). La passe précédente sert aussi à donner des repères supplémentaires (par exemple, l'étymologie apparaît toujours derrière la balise <form> ou la balise <gramGrp>). Cette passe se compose elle aussi de trois programmes, l'un pour l'étymologie, le deuxième pour les renvois, et le troisième pour les proverbes et les exemples qui utilisent la même balise TEI (à un type près).

2.2.1 Repérage de l'étymologie (balise <etym>)

Le programme permettant de récupérer les notes étymologiques du *Petit Larousse Illustré* a été l'un des plus difficiles à mettre en place, en raison de la diversité des structures que peut prendre ce type de notes. Les notes étymologiques apparaissent toujours derrière les balises <form> ou <gramGrp>, à l'intérieur d'une parenthèse. À l'intérieur de cette parenthèse, on trouve la mention d'une langue, d'un étymon en italiques et de sa glose.
Conformément aux recommandations de la TEI, les balises ont été posées de la manière suivante : la langue d'origine du mot a été balisée avec <lang>. Les langues sont repérables relativement aisément, dans la mesure où leur liste est finie. Les abréviations pour une même langue sont parfois multiples,

mais les différentes relectures ont permis d'en faire une liste exhaustive. Le nom de la langue est par ailleurs souvent précédé de « du » ou « de le », quand il ne suit pas directement la parenthèse ouvrante. Les étymons ont été balisés avec <mentioned>. Leur repérage s'est fait grâce au fait qu'ils sont en italiques. Enfin, la balise <gloss> a été utilisée pour marquer la glose ou la traduction de l'étymon, qui apparaît en général après l'étymon suivi d'une virgule.

2.2.2 Repérage des renvois (balise <xr>)

Les renvois sont de trois types différents, et nous avons une fois encore pu suivre les recommandations de la TEI pour les annoter. Ils se présentent de la manière suivante : <xr><lbl>nature du renvoi</lbl> <ref=#objet du renvoi>objet du renvoi</ref></xr>

Par *nature du renvoi*, nous entendons les choses suivantes : le renvoi est un renvoi simple, indiqué dans le dictionnaire par « v. » signifiant « voir » ; le renvoi est antonymique ou synonymique et il est indiqué dans le dictionnaire par l'abréviation « ant. » ou « syn. ».

La balise <ref> permet de faire un lien hypertexte sur l'article ayant l'objet du renvoi comme vedette.

2.2.3 Repérage des exemples et proverbes (balises <cit> et <quote>)

Les exemples et les proverbes sont balisés de façon très proche dans la TEI bien qu'il s'agisse d'objets en apparence éloignés. On les balise de la façon suivante : `<cit type = "XX"><quote> texte</quote></cit>`, avec XX prenant pour valeur soit « prov » soit « example », en fonction de la nature du texte. Les exemples sont repérables aisément car ils sont en italiques et toujours précédés de la marque typographique « : ». Les proverbes sont signalés explicitement par l'abréviation « prov » en petites capitales, de la marque « : » et sont en gras.

À la fin de la deuxième passe d'annotation, les entrées sont ainsi balisées (nous ajoutons l'entrée de « cubique » aux trois autres pour pouvoir tout illustrer) :

```
<EntryFree><form><orth>CUBIQUE</orth></form>
<gramGrp><pos>adj.</pos></gramGrp> Qui appartient au
cube <cit type=”example”>: <quote>racine cubique.</
quote></cit> (<xr type=”renvoi”><lbl>V.</lbl> <ref target
=”#racine”>racine</ref></xr>). Qui a la forme d&#39;un
cube.</EntryFree>
<EntryFree><form><orth>CUEILLEUR</orth></
form>, <form type=”infl”><orth type=”partial” mot_
entier=”cueilleuse”>EUSE</orth> <pron>(keu, ll mll.,
eu-ze)</pron></form> <gramGrp><pos>n.</pos></gramGrp>
Celui celle qui cueille. <usg type=”plev”>Peu us.</usg>)</
EntryFree>
<EntryFree><form><orth>CUILLER</orth></form> ou
<form><orth>CUILLÈRE</orth> <pron>(kui, ll mll., è-re)</
```

```
pron></form> <gramGrp><pos>n.</pos> <gen>f.</gen></
gramGrp> <etym>(<lang>lat.</lang> <mentioned>cochleare ;</
mentioned> de <mentioned type="de">cochlea</mentioned>,
<gloss>coquille</gloss>)</etym>. Ustensile de table,
composé d'un manche et d'une partie creuse pour puiser
les aliments liquides ou peu consistants. Cuillère à pot,
grande cuillère de cuisine. Ustensile servant à puiser les
métaux en fusion.</EntryFree>
<EntryFree><form><orth>CUISINIER</orth> <pron>(zi-ni-é)</
pron></form>, <form type="infl"><orth type="partial" mot_
entier="cuisinière">ÈRE</orth></form> <gramGrp><pos>n.</
pos></gramGrp> Qui fait la cuisine. <gramGrp><pos>N.</
pos> <gen>f.</gen></gramGrp> Appareil en fonte ou en tôle,
muni d'un ou de deux foyers, et à l'aide duquel on peut
faire cuire les aliments, tout en chauffant un appartement.
Sorte de rôtissoire, de coquille, destinée au grillage des
viandes.</EntryFree>
```

2.3 Troisième passe d'annotation

La troisième passe d'annotation vient clore le processus avec le balisage des éléments restants. Un premier programme vient baliser les sous-entrées et un second les définitions. Là encore, pour cette phase d'annotation sont utilisés les résultats de la précédente.

2.3.1 Pose des balises <re> pour les sous-entrées

Les balises <re> signifient « related entry ». Elles doivent être typées par la valeur « exp », « derive » ou « pron ». La première indique une expression contenant la vedette, la deuxième un dérivé morphologique, et la troisième une utilisation pronominale quand la vedette est un verbe. Les sous-entrées sont repérées de la façon suivante : les expressions sont derrière un point ou une balise fermante et apparaissent en italiques. Les dérivés ou formes pronominales apparaissent eux aussi derrière un point ou une balise fermante mais sont eux en gras.

2.3.2 Poses des balises <def> pour les définitions

Les définitions sont les éléments portant le moins de marques distinctives. Elles apparaissent en caractères droits, commencent la plupart du temps par une virgule (la première définition de l'entrée mise à part, puisqu'elle démarre derrière une indication grammaticale).

À la fin du processus d'annotation, les entrées sont donc balisées de la manière suivante :

```
<EntryFree><form><orth>CUBIQUE</orth></form>
<gramGrp><pos>adj.</pos></gramGrp> <def>Qui appartient au
cube</def> <cit type="example">: <quote>racine cubique.</
quote></cit> (<xr type="renvoi"><lbl>V.</lbl> <ref target
="#racine">racine</ref></xr>.) <def>Qui a la forme d&#39;un
```

```
cube.</def></EntryFree>
<EntryFree><form><orth>CUEILLEUR</orth></
form>, <form type="infl"><orth type="partial" mot_
entier="cueilleuse">EUSE</orth> <pron>(keu, ll mll.,
eu-ze)</pron></form> <gramGrp><pos>n.</pos></gramGrp>
<def>Celui celle qui cueille</def>. (<usg type="plev">Peu
us.</usg>)</EntryFree>
<EntryFree><form><orth>CUILLER</orth></form> ou
<form><orth>CUILLÈRE</orth> <pron>(kui, ll mll., è-re)</
pron></form> <gramGrp><pos>n.</pos> <gen>f.</gen></
gramGrp> <etym>(<lang>lat.</lang> <mentioned>cochleare ;</
mentioned> de <mentioned type="de">cochlea</mentioned>,
<gloss>coquille</gloss>)</etym>. <def>Ustensile de table,
composé d&#39;un manche et d&#39;une partie creuse pour
puiser les aliments liquides ou peu consistants</def>.
<re type='exp'><form>Cuillère à pot</form>, <def>grande
cuillère de cuisine. Ustensile servant à puiser les métaux
en fusion.</def></re></EntryFree>
<EntryFree><form><orth>CUISINIER</orth> <pron>(zi-ni-é)</
pron></form>, <form type="infl"><orth type="partial" mot_
entier="cuisinière">ÈRE</orth></form> <gramGrp><pos>n.</
pos></gramGrp> <def>Qui fait la cuisine</def>.
<gramGrp><pos>N.</pos> <gen>f.</gen></gramGrp> <def>Appareil
en fonte ou en tôle, muni d&#39;un ou de deux foyers, et à
l&#39;aide duquel on peut faire cuire les aliments, tout en
chauffant un appartement. Sorte de rôtissoire, de coquille,
destinée au grillage des viandes.</def></EntryFree>
```

2.4 Les difficultés d'une annotation automatisée

L'annotation automatisée ne s'est bien sûr par réalisée dans la facilité. En effet, sous l'apparente régularité de la mise en forme se cachent nombre d'écueils qu'il faut contourner, affronter, voire éliminer. Ne serait-ce que parce que le *Petit Larousse* est une œuvre humaine, il contient des erreurs, et ses rédacteurs ont parfois changé de méthode en cours de route. Parfois aussi, des éléments de mise en forme ont échappé à notre vigilance et ont demandé des ajustements. Aussi, le travail de balisage des fichiers a-t-il été précédé de nombreux tests, et de temps à autre d'échecs, à cause d'irrégularités aussi bien dans la mise en forme que d'irrégularités dans la structure des articles.

2.4.1 Irrégularités de mise en forme

Il y a beaucoup d'irrégularités dans la mise en forme, probablement dues à des erreurs, des oublis ou des problèmes d'impression. Ces erreurs sont souvent isolées, et elles nécessitent une correction manuelle de la balise. Certaines irrégularités en revanche sont répétées et ont été trouvées lors des relectures. En voici deux exemples :

Les abréviations annoncées ne sont pas celles utilisées : nous l'avons déjà mentionné plusieurs fois dans l'article, de nombreuses balises sont posées sur des éléments abrégés. À la relecture, on s'aperçoit que les abréviations

données par les auteurs du dictionnaire ne sont pas toujours celles qui sont réellement utilisées, et surtout qu'un même mot peut être abrégé de différentes façons.

Les prononciations sont entre crochets et non plus entre parenthèses : il a fallu une relecture du premier fichier où les prononciations étaient balisées pour que nous nous apercevions qu'une partie des prononciations n'étaient pas entre parenthèses mais entre crochets, et que de plus il ne s'agit pas d'une erreur. Pour une raison que nous comprenons mais que nous n'avions pas devinée, lorsque les entrées sont des verbes pronominaux, la vedette contient le pronom réfléchi entre parenthèses, après le verbe lui-même. Aussi, la prononciation apparaissant juste après une parenthèse fermante, les auteurs ont jugé bon (probablement pour éviter une répétition de parenthèses) de la faire apparaître entre crochets. Il a donc fallu ajouter un cas supplémentaire au programme de balisage des prononciations.

2.4.2 Irrégularités de structure

Certaines irrégularités sont bien plus difficiles à gérer que celles précédemment citées. En effet, parfois, c'est la structure même de l'information qui est bousculée ou irrégulière, et réaliser un programme d'annotation automatique devient un défi. Ainsi, le programme d'annotation des indications étymologiques a été long à réaliser et surtout, il a fallu faire de nombreuses corrections manuelles devant l'impossibilité de régler tous les problèmes dans le temps qui nous était imparti. En théorie, l'étymologie se trouve après l'information grammaticale et se structure de la façon suivante :

```
Etymologie = (nom d'une langue en abrégé + l'étymon en
italique + virgule + glose)
```

On retrouve en effet ceci dans les entrées suivantes et dans de nombreuses autres :

ABJECT (ab-jèkt'), E adj. (lat. abjectus, jeté hors.)

ABJURATION (ra-si-on) n. f. (lat. abjuratio, reniement.)

ABNÉGATION (si-on) n. f. (lat. abnegatio, action de nier.)

En pratique, on trouve une multitude de configurations qu'il n'est pas facile de décrire et de généraliser. Ainsi, on relève les situations suivantes :

ABLÉGAT (ga) n. m. (préf. ab, et lat. legatus, envoyé.)

Ici est offerte l'indication de construction morphologique, mais d'où vient le préfixe « ab » ? Est-il français ou est il latin ? De plus, le tout est suivi de la mention d'une racine latine et de sa traduction.

ABOLIR v. a. (lat. abolere.) : ici, on ne trouve pas de traduction de l'étymon.

ABOI n. m. (de aboyer.) Ici, on ne trouve pas une racine latine, mais l'indication du verbe ayant servi à la dérivation.

ABÎME n. m. (du gr. a priv., et bussos, fond.) Ici, le mot est composé de deux étymons, ce que le programme traitait parfaitement, mais le terme « priv. » n'est pas une traduction, mais l'abréviation d'une glose.

ABÉE (bé) n. f. (du vx fr. bée, auj. baie, ouverture.) : Ici, est donnée une étymologie, avec l'indication de l'évolution du terme en français contemporain.

ABRÉGER (jé) v. a. (lat. abbreviare ; de brevis, court. - Prend un e ouvert devant une syllabe muette : […])

Ici, l'étymologie est latine pour le mot, tout en étant donné aussi pour la morphologie latine. Elle est par ailleurs suivie d'indications de conjugaison fréquentes dans le *Petit Larousse* et que notre programme avait prévu de gérer dès le départ.

Nous pouvons donc constater que les cas étaient nombreux, et qu'automatiser le repérage de toutes ces configurations était difficile. Dans certaines situations, il a été possible d'adapter le script. Ainsi, nous avons pu tenir compte des cas suivants :

ABÉE (bé) n. f. (du vx fr. bée, auj. baie, ouverture.)
ABÎME n. m. (du gr. a priv., et bussos, fond.)
ABOLIR v. a. (lat. abolere.)
ABRÉGER (jé) v. a. (lat. abbreviare ; de brevis, court. - Prend un e ouvert devant une syllabe…

Pour d'autres cas, il nous a fallu ajouter un type aux balises de la TEI. Ainsi, dans les cas où le dictionnaire ne présente pas réellement une étymologie (ou pas seulement), nous avons typé la balise <etym> avec un type = « morpho » qui permet de distinguer une étymologie pure d'un découpage morphologique

ABLÉGAT (ga) n. m. (préf. ab, et lat. lega
tus, envoyé.)
ABOI n. m. (de aboyer.)

Ces entrées ont donc été balisées de la façon suivante :

```
ABÉE […] <etym>(du <lang>vx fr.</lang> <mentioned>bée</
mentioned>, <usg type="time">auj.</usg> <mentioned>baie</
mentioned>, <gloss>ouverture</gloss>)</etym>.
ABÎME(…) <etym>(du <lang>gr.</lang> <mentioned>a</
mentioned> <gloss>priv.,</gloss> et <mentioned>bussos</
mentioned>, <gloss>fond</gloss>)</etym>.
<EntryFree><form><orth>
ABRÉGER</orth> <pron>(jé)</pron></form> <gramGrp><pos>v.
a.</pos></gramGrp> <etym>(<lang>lat.</lang>
<mentioned>abbreviare ;</mentioned> de <mentioned
type="de">brevis</mentioned>, <gloss>court</gloss></etym>.
— <note type="conjugaison">Prend un e ouvert devant une
syllabe muette (…) </note>
ABOLIR</orth></form> <gramGrp><pos>v. a.</pos></gramGrp>
<etym>(<lang>lat.</lang> <mentioned>abolere</mentioned>)</
etym>.
ABLÉGAT (…) <etym type=morpho>(préf. <mentioned>ab</
mentioned>, et <lang>lat.</lang> <mentioned>legatus</
mentioned>, <gloss>envoyé</gloss>)</etym>.
ABOI (…) <etym type=morpho> (de <i>aboyer</i>) </etym>.
```

Nous avons donc montré dans cette section que le balisage automatique du dictionnaire n'a pas représenté une tâche aisée, mais il a été rendu possible

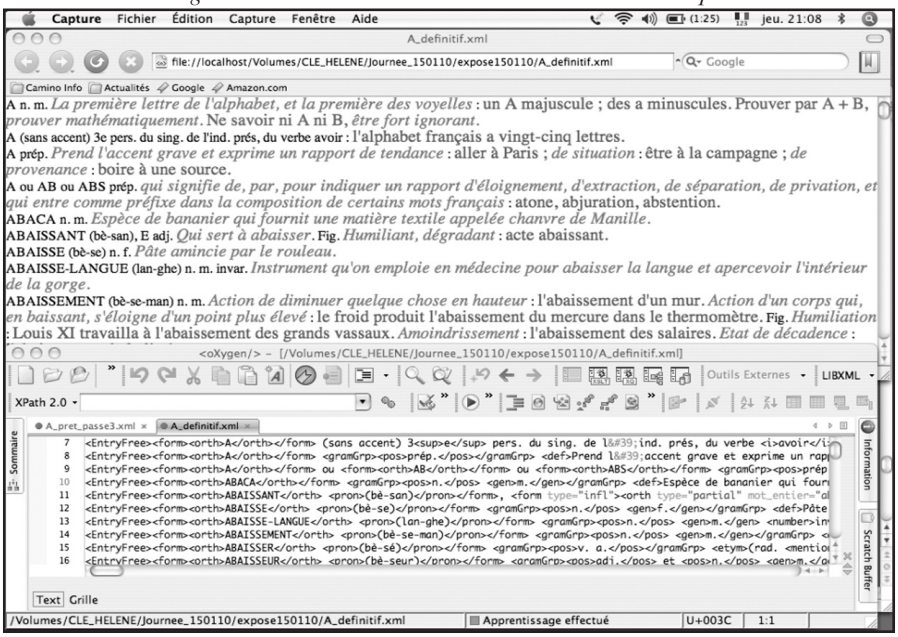
grâce à beaucoup de temps passé à observer et analyser les données. Nous avons aussi montré comment il était possible de respecter les recommandations de la TEI sans dénaturer le sens du texte et en restant précis dans le balisage sémantique et structurel du dictionnaire.
Cependant, le balisage a beau avoir été automatisé, il a fallu de nombreuses relectures humaines pour arriver au résultat que nous signalons aujourd'hui. Ces relectures auraient été encore plus difficiles si nous n'avions pas bénéficié d'outils d'aide à la relecture que nous présentons maintenant, ainsi que d'outils permettant une mise en ligne rapide et progressive des fichiers.

2.5 LES OUTILS D'AIDE À LA RELECTURE ET À LA MISE EN LIGNE

2.5.1 La relecture facilitée

Chaque passe de balisage a été relue. Il a donc fallu rendre les relectures plus efficaces et moins fastidieuses, et imaginer des méthodes de relecture simplifiée. En d'autres termes, il fallait faire ressortir les éléments récemment balisés du reste du texte du dictionnaire et faire apparaître le moins possible de balises XML pour ne pas « brouiller » la vue de la relectrice. La solution s'est donc imposée : de simples feuilles de styles. Les relectrices chargeaient le texte à relire dans un navigateur et dans un éditeur XML. Elles relisaient le texte dans le navigateur, où les éléments récemment balisés apparaissaient en couleur. Si elles notaient un oubli, ou un élément balisé à tort, elles pouvaient corriger directement dans l'éditeur XML.

Figure 1 : Écran de relecture de l'annotation automatique



2.5.2 L'interface de mise en ligne des images et du texte

De la même façon, la mise en ligne des fichiers se faisant au fur et à mesure de l'avancée des travaux et la mise en ligne des images (fort nombreuses) s'effectuant à part de celle du texte, il a fallu concevoir une interface d'administration du site confortable et permettant la mise à jour régulière du site par toute personne travaillant dessus. L'interface se présente donc ainsi :

Figure 2 : Écran d'administration de la base

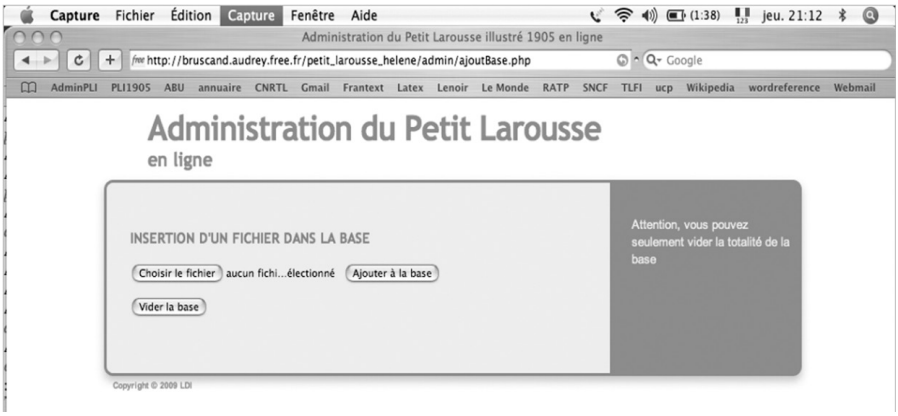
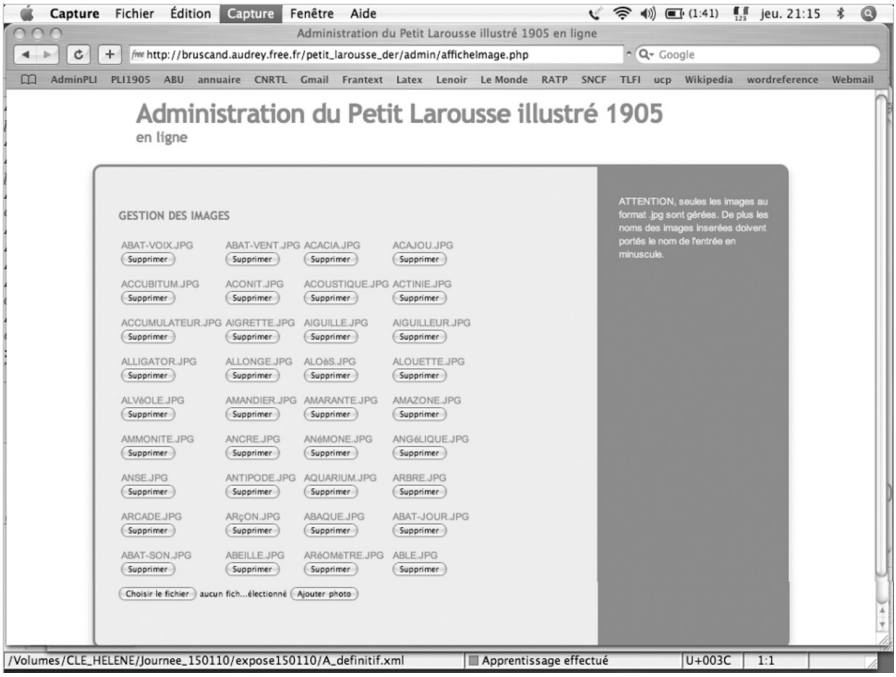


Figure 3 : Écran d'administration de la base d'images



Il suffit de simples clics pour charger les fichiers. Il n'est donc pas nécessaire d'avoir un seul administrateur du site ou un client ftp, un navigateur suffit à la mise en ligne.

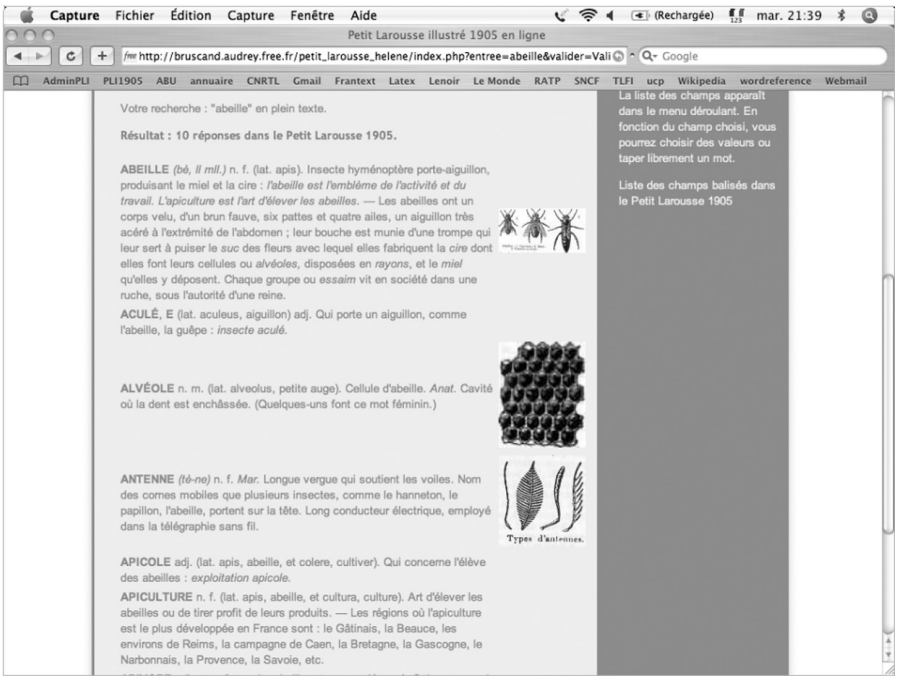
3 PRÉSENTATION DE LA BASE

Nous souhaitons ici montrer le fonctionnement de la base de données dans son ensemble, de façon à ce que les balises présentées dans les sections précédentes prennent leur sens. Le travail étant en cours d'achèvement, l'adresse qui apparaît sur les copies d'écran et certaines parties de l'interface sont provisoires. De plus, les recherches présentées ne portent que sur des mots commençant par A, B ou C, puisqu'à l'heure où nous écrivons l'article, seules ces trois lettres (représentant à elles seules le quart du dictionnaire) sont totalement relues et vérifiées.

3.1 La recherche plein texte

La méthode la plus simple pour faire une recherche dans un document électronique reste la recherche plein texte. Cette recherche s'effectue dans le texte intégral, sans tenir compte du balisage. Ainsi, elle permet de travailler sur un terme dans sa globalité, et non pas dans un contexte spécifique. La recherche plein texte dans le *Petit Larousse Illustré* de 1905 en ligne affiche l'intégralité de l'entrée où le mot recherché apparaît, ainsi que l'illustration associée. Dans l'illustration ci-dessous, on bénéficie d'une partie des résultats donnés pour la recherche plein texte du mot « abeille » :

Figure 4 : recherche plein texte



3.2 La recherche avancée

La recherche avancée a lieu sur les différents champs balisés décrits dans la section précédente. Ces champs apparaissent dans un menu déroulant, comme en témoigne l'illustration suivante :

Figure 5 : recherche avancée

3.2.1 La recherche sur une vedette ou sa forme fléchie

La recherche sur la vedette est bien entendu possible, étant donné ce que nous avons expliqué dans les paragraphes précédents. Comme le révèle la figure 6 (page 468), la recherche sur la vedette fait elle aussi, à la manière d'une recherche plein texte, apparaître l'illustration associée à l'article.
La recherche sur la vedette permet aussi de procéder à une recherche sur la forme fléchie des noms et des adjectifs. Ainsi, la recherche faite sur la vedette « cireuse » donnera le résultat de la figure 7 (page 468).

3.2.1 La recherche sur une note étymologique

Nous proposons trois entrées pour la recherche sur les notes étymologiques. La première sur la langue d'origine des mots, la deuxième sur l'étymon lui-même, et la troisième sur le sens de l'étymon. Aussi, après avoir cliqué sur le premier menu déroulant, l'utilisateur en voit apparaître un second qui lui permet de choisir entre les trois possibilités, comme on le constate dans la capture d'écran page 469 (figure 8).

Figure 6 : Recherche sur la vedette

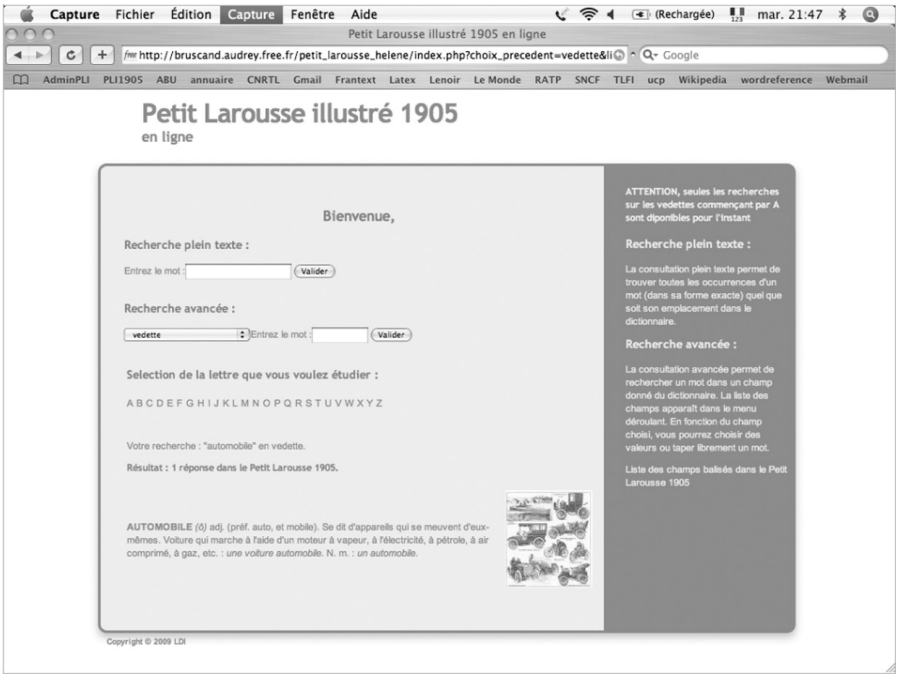


Figure 7 : Recherche sur la forme fléchie d'une vedette

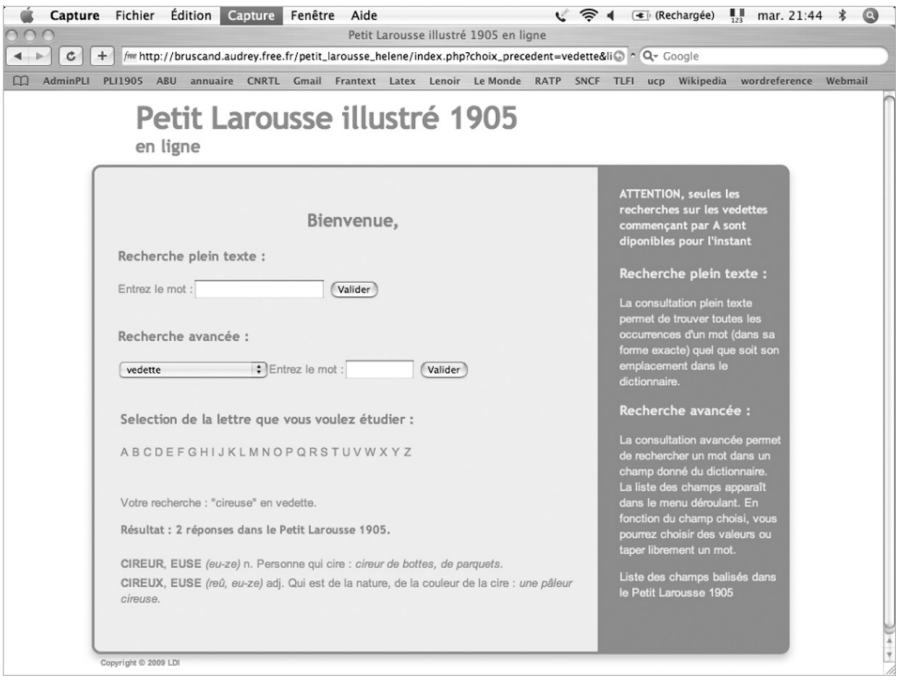


Figure 8 : Recherche sur les notes étymologiques

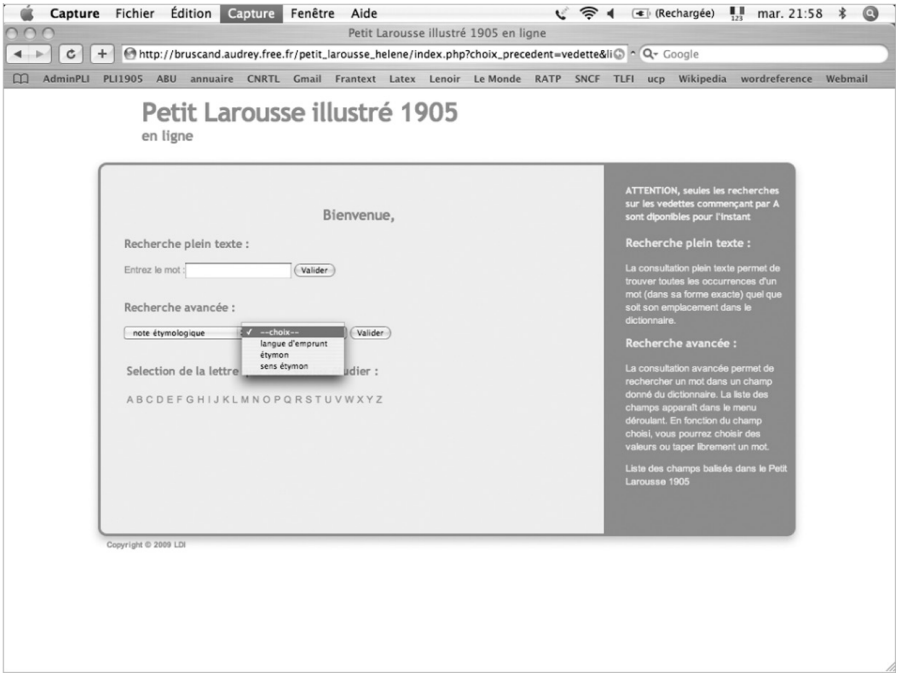


Figure 9 : Recherche sur la langue d'origine

Figure 10 : Recherche sur un étymon

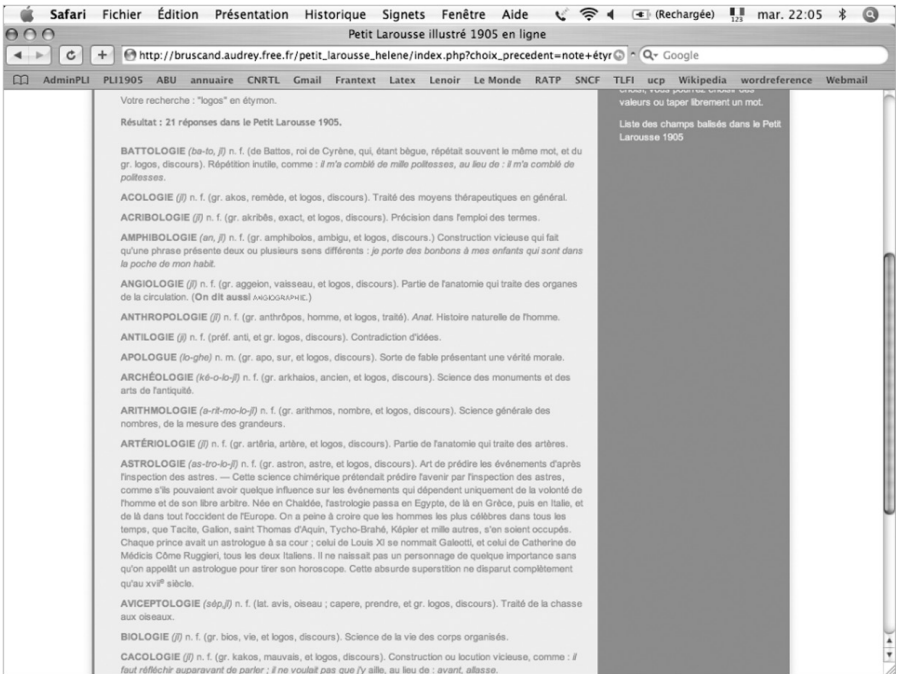
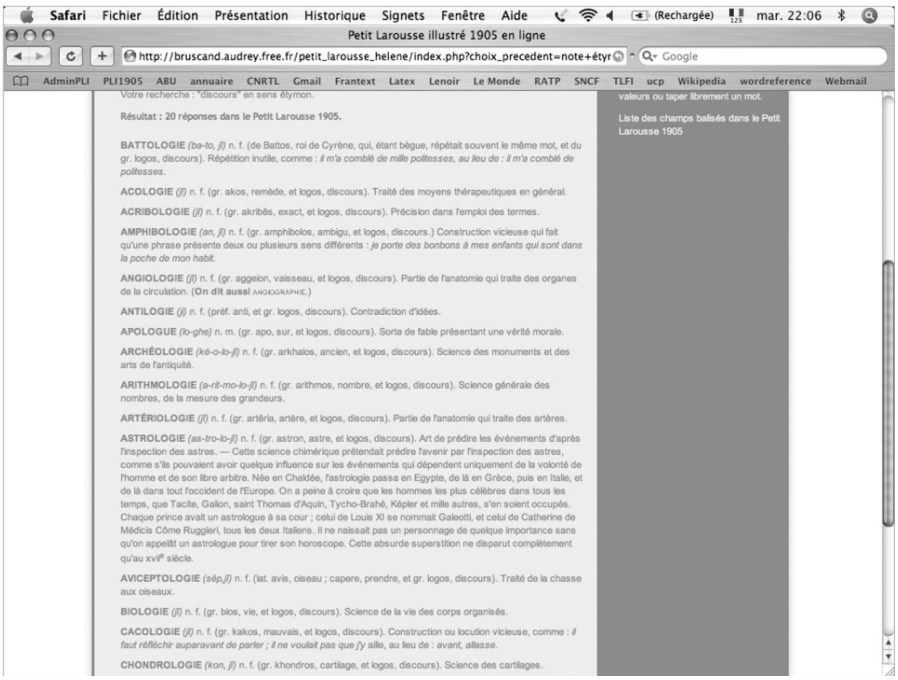


Figure 11 : Recherche sur la traduction de l'étymon



Dans les trois captures d'écran 9, 10 et 11, nous présentons le résultat des recherches suivantes : la recherche de tous les mots provenant de l'ancien haut allemand ; celle correspondant à tous les mots ayant pour étymologie le mot grec « logos » ; enfin, la recherche de tous les mots dont le sens de l'étymon est « discours ». On notera que pour ces deux dernières recherches les résultats ne sont pas parallèles, *logos* étant la plupart du temps traduit par « discours » mais parfois aussi par « traité ».

3.2.3 La recherche sur les renvois

La recherche sur les renvois de tous types a ceci d'intéressant qu'elle propose à l'utilisateur un lien hypertexte sur la cible du renvoi. Ainsi, dans l'exemple suivant, l'utilisateur peut-il cliquer sur « abhorrer » pour accéder à la définition de ce mot.

Figure 12 : recherche sur les renvois

3.2.4 La distinction définition lexicale/définition encyclopédique

Nous avons choisi de distinguer les définitions lexicales des définitions encyclopédiques. En effet, bien que cette distinction ne soit pas toujours évidente à réaliser sur le fond, il nous a paru important de respecter le fait que les auteurs du *Petit Larousse illustré* de 1905 l'ont faite sur la forme. En effet, à la fin de certains articles du dictionnaire, un texte relativement long (en tout cas comparé au texte de la définition) apparaît après un tiret long. Aussi, nous avons décidé de conserver cette distinction, même s'il faut l'avouer, souvent,

Figure 13 : Résultats de la recherche du mot « poisson » dans une définition

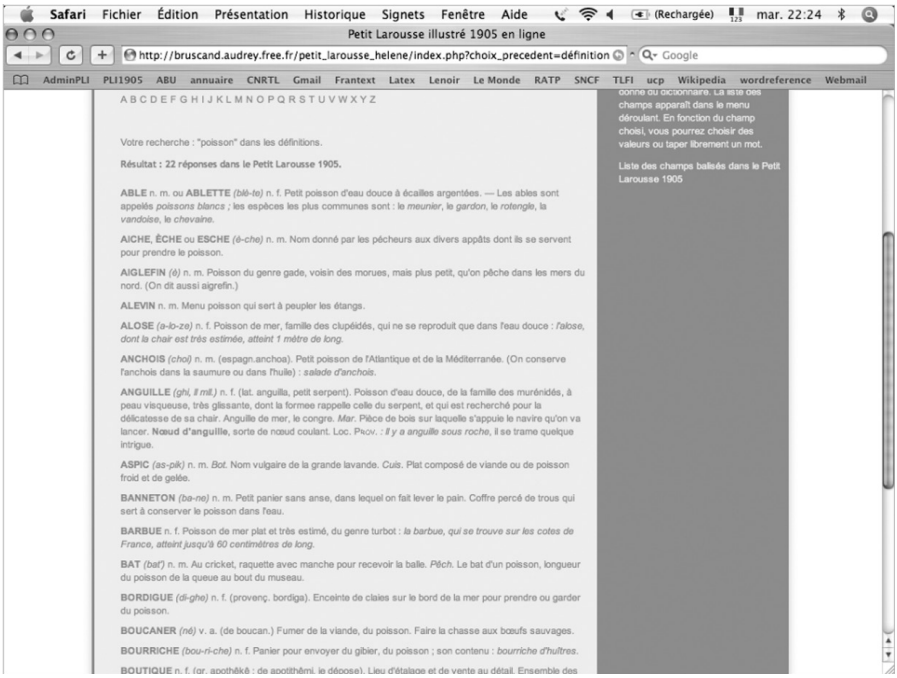
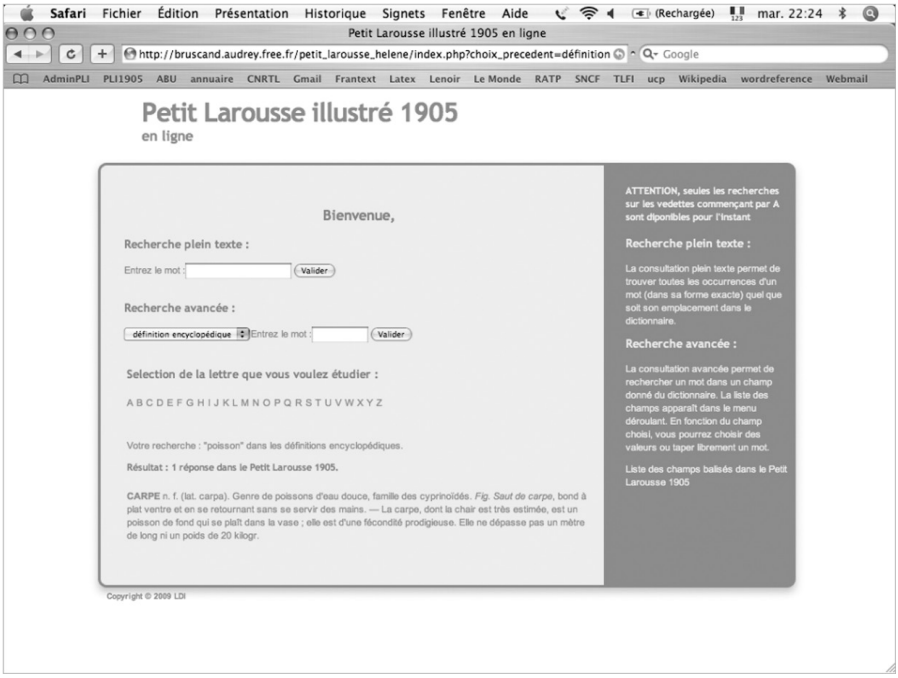


Figure 14 : Résultats de la recherche du mot « poisson » dans une définition encyclopédique

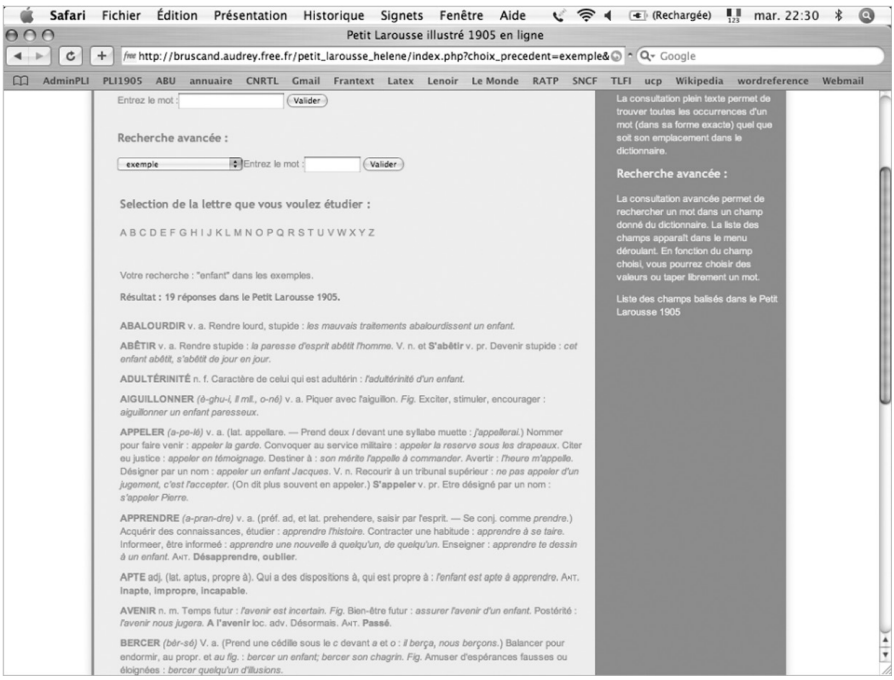


les définitions données sous la forme de définitions « lexicales » revêtent un fort caractère encyclopédique.
Ainsi, la recherche sur le mot « poisson » donne les résultats suivants pour, dans l'ordre, une recherche dans les définitions (figure 13, page 472) puis une recherche dans les définitions encyclopédiques (figure 14).

3.2.5 La recherche sur les exemples

La recherche portant sur les exemples est probablement la plus intéressante à mener sur un plan extralinguistique, presque sociologique. En effet, les exemples du *Petit Larousse Illustré* étant forgés, ils constituent de façon très claire le reflet d'une époque et d'une idéologie. Aussi, une recherche tout à fait significative pourrait-elle être effectuée sur les exemples contenant le mot « enfant ». Les commentaires sociologiques sont inutiles : notre lecteur arrivera aux mêmes conclusions que nous à la lecture des résultats présentés ci-dessous :

Figure 15 : Résultat de la recherche du mot « enfant » dans les exemples



4 CONCLUSIONS

L'équipe du laboratoire LDI est donc en mesure de présenter une nouvelle base de données lexicales importante et ce à plusieurs points de vue.
Il s'agit tout d'abord d'une base de données lexicales importante du point de vue historique, puisqu'elle correspond au premier dictionnaire illustré démocratique de l'histoire éditoriale francophone. Il était donc fondamental pour la métalexicographie de disposer d'une telle base.

Ensuite, cette base offre des possibilités de recherche avancée qui en permettent une interrogation fine et l'extraction de résultats très intéressants, qu'il serait certes possible d'obtenir à partir d'une version papier, mais au prix d'efforts infiniment plus contraignants.

Enfin, le *Petit Larousse Illustré* de 1905 en ligne représente un fait informatique important. Bien que rédigé au début du vingtième siècle, alors que ses auteurs, probablement visionnaires, ignoraient tout de ce qu'allait être l'informatique, le dictionnaire a pu être balisé selon des standards précis. Il a certes parfois fallu inventer quelques attributs, faire quelques concessions, mais l'intégralité du dictionnaire est entrée dans les « cases » établies par les linguistes-informaticiens de la fin du vingtième siècle. Aussi, ce travail doit-il être un encouragement pour les lexicographes contemporains : il est possible de baliser un texte sans pour autant être enfermé dans un carcan.

Hélène MANUÉLIAN, Audrey BRUSCAND,
Université de Cergy-Pontoise,
Lexiques Dictionnaires Informatique (LDI, CNRS, UMR 71 87)
Nicole CHOLEWKA, Anne-Marie HETZEL
(LDI, CNRS, UMR 71 87)

BIBLIOGRAPHIE

ADRESSE DU SITE : http://dictionnaire1905.u-cergy.fr

CARON, P., DAGENAIS, L., GONFROY, G. 1996. Le programme d'informatisation du « Dictionaire critique de la langue française » de l'abbé Jean-François Féraud (1787), CHWP B.6, publ. mai 1996.

DENDIEN J., PIERREL, J-M. 2003. « Le trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence », dans *Traitement automatique des langues*, vol. 44, n° 2.

IDE, N., VÉRONIS, J. 1994. Chapter 12 : Print dictionaries. In C. M. Sperberg-McQueen & L. Burnard (Eds.), Guidelines for Electronic Text Encoding and Interchange (pp. 321-370). Chicago, Oxford : The Text Encoding Initiative.

PRUVOST, J. 2000. *Dictionnaires et nouvelles technologies*, Presses universitaires de France.

PRUVOST, J.2004. *La dent-de-lion, la semeuse et le Petit Larousse, Larousse*, Paris.

TEI consortium, *Print Dictionaries*, TEI P5, http://www.tei-c.org/release/doc/tei-p5-doc/html/DI.html